July 2008

Grassmann Discriminant Analysis: a Unifying View on Subspace-Based Learning

Jihun Hamm
University of Pennsylvania

Daniel D. Lee
University of Pennsylvania, ddlee@seas.upenn.edu

Reprinted from the Proceedings of the 25th International Conference on Machine Learning (ICML 2008), July 2008.
URL: http://icml2008.cs.helsinki.fi/index.shtml

Grassmann Discriminant Analysis: a Unifying View on Subspace-Based Learning

Abstract

In this paper we propose a discriminant learning framework for problems in which data consist of linear subspaces instead of vectors. By treating subspaces as basic elements, we can make learning algorithms adapt naturally to the problems with linear invariant structures. We propose a unifying view on the subspace-based learning method by formulating the problems on the Grassmann manifold, which is the set of fixed-dimensional linear subspaces of a Euclidean space. Previous methods on the problem typically adopt an inconsistent strategy: feature extraction is performed in the *Euclidean* space while *non-Euclidean* distances are used. In our approach, we treat each subspace as a point in the Grassmann space, and perform feature extraction and classification in the same space. We show feasibility of the approach by using the Grassmann kernel functions such as the Projection kernel and the Binet-Cauchy kernel. Experiments with real image databases show that the proposed method performs well compared with state-of- the-art algorithms.

Comments

Reprinted from the Proceedings of the 25th International Conference on Machine Learning (ICML 2008), July 2008.

URL: http://icml2008.cs.helsinki.fi/index.shtml

Grassmann Discriminant Analysis: a Unifying View on Subspace-Based Learning

Jihun Hamm JHHAM@SEAS.UPENN.EDU
Daniel D. Lee DDLEE@SEAS.UPENN.EDU
GRASP Laboratory, University of Pennsylvania, Philadelphia, PA 19104 USA

Abstract

In this paper we propose a discriminant learning framework for problems in which data consist of linear subspaces instead of vectors. By treating subspaces as basic elements, we can make learning algorithms adapt naturally to the problems with linear invariant structures. We propose a unifying view on the subspace-based learning method by formulating the problems on the Grassmann manifold, which is the set of fixed-dimensional linear subspaces of a Euclidean space. Previous methods on the problem typically adopt an inconsistent strategy: feature extraction is performed in the *Euclidean* space while *non-Euclidean* distances are used. In our approach, we treat each subspace as a point in the Grassmann space, and perform feature extraction and classification in the same space. We show feasibility of the approach by using the Grassmann kernel functions such as the Projection kernel and the Binet-Cauchy kernel. Experiments with real image databases show that the proposed method performs well compared with state-of-the-art algorithms.

1. Introduction

We often encounter learning problems in which the basic elements of the data are *sets of vectors* instead of vectors. Suppose we want to recognize a person from multiple pictures of the individual, taken from different angles, under different illumination or at different places. When comparing such sets of image vectors, we are free to define the similarity between sets based on the similarity between image vectors (Shakhnarovich et al., 2002; Kondor & Jebara, 2003; Zhou & Chellappa, 2006).

In this paper, we specifically focus on those data that can be modeled as a collection of linear subspaces. In the example above, let's assume that the set of images of a single person is well approximated by a low dimensional subspace (Turk & Pentland, 1991), and the whole data is the collection of such subspaces. The benefits of using subspaces are two-fold: 1) comparing two subspaces is cheaper than comparing two sets directly when those sets are very large, and 2) it is more robust to missing data since the subspace can 'fill-in' the missing pictures. However the advantages come with the challenge of representing and handling the subspaces appropriately.

We approach the subspace-based learning problems by formulating the problems on the Grassmann manifold, the set of fixed-dimensional linear subspaces of a Euclidean space. With this unifying framework we can make analytic comparisons of the various distances of subspaces. In particular, we single out those distances that are induced from the *Grassmann kernels*, which are positive definite kernel functions on the Grassmann space. The Grassmann kernels allow us to use the usual kernel-based algorithms on this unconventional space and to avoid ad hoc approaches to the problem.

We demonstrate the proposed framework by using the Projection metric and the Binet-Cauchy metric and by applying kernel Linear Discriminant Analysis to classification problems with real image databases.

1.1. Contributions of the Paper

Although the Projection metric and the Binet-Cauchy metric were previously used (Chang et al., 2006; Wolf & Shashua, 2003), their potential for subspace-based learning has not been fully explored. In this work, we provide an analytic exposition of the two metrics as examples of the Grassmann kernels, and contrast the

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

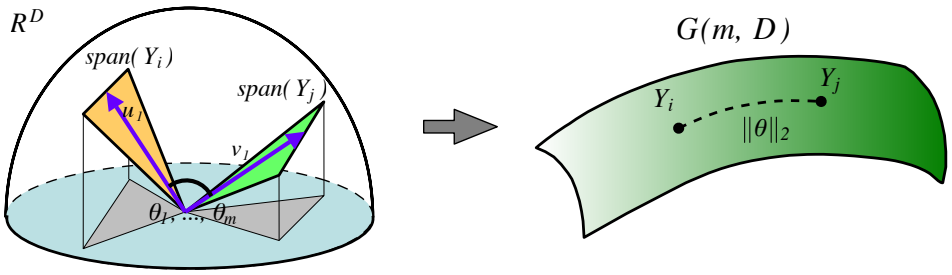

Figure 1. Principal angles and Grassmann distances. Let span(Y_i) and span(Y_j) be two subspaces in the Euclidean space $\mathbb{R}^D$ on the left. The distance between two subspaces span(Y_i) and span(Y_j) can be measured by the principal angles $\boldsymbol{\theta} = [\theta_1, \dots, \theta_m]'$ using the usual innerproduct of vectors. In the Grassmann manifold viewpoint, the subspaces span(Y_i) and span(Y_j) are considered as two points on the manifold $\mathcal{G}(m, D)$, whose Riemannian distance is related to the principal angles by $d(Y_i, Y_j) = \|\boldsymbol{\theta}\|_2$. Various distances can be defined based on the principal angles.

two metrics with other metrics used in the literature.

Several subspace-based classification methods have been previously proposed (Yamaguchi et al., 1998; Sakano, 2000; Fukui & Yamaguchi, 2003; Kim et al., 2007). However, these methods adopt an inconsistent strategy: feature extraction is performed in the *Euclidean* space when *non-Euclidean* distances are used. This inconsistency can result in complications and weak guarantees. In our approach, the feature extraction and the distance measurement are integrated around the Grassmann kernel, resulting in a simpler and better-understood formulation.

The rest of the paper is organized as follows. In Sec. 2 and 3 we introduce the Grassmann manifolds and derive various distances on the space. In Sec. 4 we present a kernel view of the problem and emphasize the advantages of using positive definite metrics. In Sec. 5 we propose the Grassmann Discriminant Analysis and compare it with other subspace-based discrimination methods. In Sec. 6 we test the proposed algorithm for face recognition and object categorization tasks. We conclude in Sec. 7 with a discussion.

2. Grassmann Manifold and Principal Angles

In this section we briefly review the Grassmann manifold and the principal angles.

Definition 1 *The Grassmann manifold $\mathcal{G}(m, D)$ is the set of m-dimensional linear subspaces of the $\mathbb{R}^D$.*

The $\mathcal{G}(m, D)$ is a $m(D-m)$-dimensional compact Riemannian manifold.[1] An element of $\mathcal{G}(m, D)$ can be represented by an orthonormal matrix Y of size D by m such that $Y'Y = I_m$, where I_m is the m by m identity matrix. For example, Y can be the m basis vectors of a set of pictures in $\mathbb{R}^D$. However, the matrix representation of a point in $\mathcal{G}(m, D)$ is not unique: two matrices Y_1 and Y_2 are considered the same if and only if span(Y_1) = span(Y_2), where span(Y) denotes the subspace spanned by the column vectors of Y. Equivalently, span(Y_1) = span(Y_2) if and only if $Y_1 R_1 = Y_2 R_2$ for some $R_1, R_2 \in \mathcal{O}(m)$. With this understanding, we will often use the notation Y when we actually mean its equivalence class span(Y), and use $Y_1 = Y_2$ when we mean span(Y_1) = span(Y_2), for simplicity.

Formally, the Riemannian distance between two subspaces is the length of the shortest geodesic connecting the two points on the Grassmann manifold. However, there is a more intuitive and computationally efficient way of defining the distances using the *principal angles* (Golub & Loan, 1996).

Definition 2 *Let Y_1 and Y_2 be two orthonormal matrices of size D by m. The principal angles $0 \le \theta_1 \le \cdots \le \theta_m \le \pi/2$ between two subspaces* span(Y_1) *and* span(Y_2)*, are defined recursively by*

$$\cos\theta_k = \max_{\boldsymbol{u_k} \in \mathrm{span}(Y_1)} \max_{\boldsymbol{v_k} \in \mathrm{span}(Y_2)} \boldsymbol{u_k}'\boldsymbol{v_k}, \quad \textit{subject to}$$

$$\boldsymbol{u_k}'\boldsymbol{u_k} = 1,\ \boldsymbol{v_k}'\boldsymbol{v_k} = 1,$$

$$\boldsymbol{u_k}'\boldsymbol{u_i} = 0,\ \boldsymbol{v_k}'\boldsymbol{v_i} = 0,\ (i = 1, \dots, k-1).$$

In other words, the first principal angle θ_1 is the smallest angle between all pairs of unit vectors in the first and the second subspaces. The rest of the principal

[1] $\mathcal{G}(m, D)$ can be derived as a quotient space of orthogonal groups $\mathcal{G}(m, D) = \mathcal{O}(D)/\mathcal{O}(m) \times \mathcal{O}(D-m)$, where $\mathcal{O}(m)$ is the group of m by m orthonormal matrices. We refer the readers to (Wong, 1967; Absil et al., 2004) for details on the Riemannian geometry of the space.

angles are similarly defined. It is known (Wong, 1967; Edelman et al., 1999) that the principal angles are related to the geodesic distance by $d_G^2(Y_1, Y_2) = \sum_i \theta_i^2$ (refer to Fig. 1.)

The principal angles can be computed from the Singular Value Decomposition (SVD) of $Y_1'Y_2$,

$$Y_1'Y_2 = U(\cos\Theta)V', \tag{1}$$

where $U = [\boldsymbol{u_1}\ ...\ \boldsymbol{u_m}]$, $V = [\boldsymbol{v_1}\ ...\ \boldsymbol{v_m}]$, and $\cos\Theta$ is the diagonal matrix $\cos\Theta = \mathrm{diag}(\cos\theta_1\ ...\ \cos\theta_m)$. The cosines of the principal angles $\cos\theta_1,\ ...\ , \cos\theta_m$ are also known as *canonical correlations*.

Although the definition can be extended to the cases where Y_1 and Y_2 have different number of columns, we will assume Y_1 and Y_2 have the same size D by m throughout this paper. Also, we will occasionally use $\mathcal{G}$ instead of $\mathcal{G}(m, D)$ for simplicity.

3. Distances for Subspaces

In this paper we use the term *distance* as any assignment of nonnegative values for each pair of points in a space $\mathcal{X}$. A valid *metric* is, however, a distance that satisfies the additional axioms:

Definition 3 *A real-valued function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a metric if*

1. $d(x_1, x_2) \geq 0$,
2. $d(x_1, x_2) = 0$ *if and only if* $x_1 = x_2$,
3. $d(x_1, x_2) = d(x_2, x_1)$,
4. $d(x_1, x_2) + d(x_2, x_3) \leq d(x_1, x_3)$,

for all $x_1, x_2, x_3 \in \mathcal{X}$.

A distance (or a metric) between subspaces $d(Y_1, Y_2)$ has to be invariant under different representations $d(Y_1, Y_2) = d(Y_1R_1, Y_2R_2),\ \ \forall R_1, R_2 \in \mathcal{O}(m)$.

In this section we introduce various distances for subspaces derivable from the principal angles.

3.1. Projection Metric and Binet-Cauchy Metric

We first underline two main distances of this paper.

1. *Projection metric*

$$d_P(Y_1, Y_2) = \left(\sum_{i=1}^{m} \sin^2\theta_i\right)^{1/2} = \left(m - \sum_{i=1}^{m} \cos^2\theta_i\right)^{1/2}. \tag{2}$$

The Projection metric is the 2-norm of the sine of principal angles (Edelman et al., 1999; Wang et al., 2006).

2. *Binet-Cauchy metric*

$$d_{BC}(Y_1, Y_2) = \left(1 - \prod_i \cos^2\theta_i\right)^{1/2}. \tag{3}$$

The Binet-Cauchy metric is defined with the product of canonical correlations (Wolf & Shashua, 2003; Vishwanathan & Smola, 2004).

As the names hint, these two distances are in fact valid metrics satisfying Def. 3. The proofs are deferred until Sec. 4.

3.2. Other Distances in the Literature

We describe a few other distances used in the literature. The principal angles are the keys that relate these distances.

1. *Max Correlation*

$$d_{\mathrm{Max}}(Y_1, Y_2) = \left(1 - \cos^2\theta_1\right)^{1/2} = \sin\theta_1. \tag{4}$$

The max correlation is a distance based on only the largest canonical correlation $\cos\theta_1$ (or the smallest principal angle θ_1). This max correlation was used in previous works (Yamaguchi et al., 1998; Sakano, 2000; Fukui & Yamaguchi, 2003).

2. *Min Correlation*

$$d_{\mathrm{Min}}(Y_1, Y_2) = \left(1 - \cos^2\theta_m\right)^{1/2} = \sin\theta_m. \tag{5}$$

The min correlation is defined similarly to the max correlation. However, the min correlation is more closely related to the Projection metric: we can rewrite the Projection metric as $d_P = 2^{-1/2}\ \|Y_1Y_1' - Y_2Y_2'\|_F$ and the min correlation as $d_{\mathrm{Min}} = \|Y_1Y_1' - Y_2Y_2'\|_2$.

3. *Procrustes metric*

$$d_{CF}(Y_1, Y_2) = 2\left(\sum_{i=1}^{m} \sin^2(\theta_i/2)\right)^{1/2}. \tag{6}$$

The Procrustes metric is the minimum distance between different representations of two subspaces $\mathrm{span}(Y_1)$ and $\mathrm{span}(Y_2)$: (Chikuse, 2003)

$$d_{CF} = \min_{R_1, R_2 \in \mathcal{O}(m)} \|Y_1R_1 - Y_2R_2\|_F = \|Y_1U - Y_2V\|_F,$$

where U and V are from (1). By definition, the distance is invariant of the choice of the

bases of $\text{span}(Y_1)$ and $\text{span}(Y_2)$. The Procrustes metric is also called chordal distance (Edelman et al., 1999). We can similarly define the minimum distance using other matrix norms such as $d_{C2}(Y_1, Y_2) = \|Y_1U - Y_2V\|_2 = 2\sin(\theta_m/2)$.

3.3. Which Distance to Use?

The choice of the best distance for a classification task depends on a few factors. The first factor is the distribution of data. Since the distances are defined with particular combinations of the principal angles, the best distance depends highly on the probability distribution of the principal angles of the given data. For example, d_{Max} uses the smallest principal angle θ_1 only, and may be robust when the data are noisy. On the other hand, when all subspaces are sharply concentrated on one point, d_{Max} will be close to zero for most of the data. In this case, d_{Min} may be more discriminative. The Projection metric d_P, which uses all the principal angles, will show intermediate characteristics between the two distances. Similar arguments can be made for the Procrustes metrics d_{CF} and d_{C2}, which use all angles and the largest angle only, respectively.

The second criterion for choosing the distance, is the degree of structure in the distance. Without any structure a distance can be used only with a simple K-Nearest Neighbor (K-NN) algorithm for classification. When a distance have an extra structure such as triangle inequality, for example, we can speed up the nearest neighbor searches by estimating lower and upper limits of unknown distances (Faragó et al., 1993). From this point of view, the max correlation is not a metric and may not be used with more sophisticated algorithms. On the other hand, the Min Correlation and the Procrustes metrics are valid metrics[2].

The most structured metrics are those which are induced from a positive definite kernel. Among the metrics mentioned so far, only the Projection metric and the Binet-Cauchy metric belong to this class. The proof and the consequences of positive definiteness are the main topics of the next section.

4. Kernel Functions for Subspaces

We have defined a valid metric on Grassmann manifolds. The next question is whether we can define a kernel function compatible with the metric. For this purpose let's recall a few definitions. Let $\mathcal{X}$ be any set, and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric real-valued function $k(x_i, x_j) = k(x_j, x_i)$ for all $x_i, x_j \in \mathcal{X}$.

Definition 4 *A real symmetric function is a (resp. conditionally) positive definite kernel function, if* $\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0$, *for all* $x_1, ..., x_n (x_i \in \mathcal{X})$ *and* $c_1, ..., c_n (c_i \in \mathbb{R})$ *for any* $n \in \mathbb{N}$. *(resp. for all* $c_1, ..., c_n (c_i \in \mathbb{R})$ *such that* $\sum_{i=1}^n c_i = 0$.*)*

In this paper we are interested in the kernel functions on the Grassmann space.

Definition 5 *A Grassmann kernel function is a positive definite kernel function on* $\mathcal{G}$.

In the following we show that the Projection metric and the Binet-Cauchy are induced from the Grassmann kernels.

4.1. Projection Metric

The Projection metric can be understood by associating a point $\text{span}(Y) \in \mathcal{G}$ with its projection matrix YY' by an embedding:

$$\Psi_P : \mathcal{G}(m, D) \to \mathbb{R}^{D \times D}, \quad \text{span}(Y) \mapsto YY'. \tag{7}$$

The image $\Psi_P(\mathcal{G}(m, D))$ is the set of rank-m orthogonal projection matrices. This map is in fact an isometric embedding (Chikuse, 2003) and the projection metric is simply a Euclidean distance in $\mathbb{R}^{D \times D}$. The corresponding innerproduct of the space is $\text{tr}\left[(Y_1Y_1')(Y_2Y_2')\right] = \|Y_1'Y_2\|_F^2$, and therefore

Proposition 1 *The Projection kernel*

$$k_P(Y_1, Y_2) = \|Y_1'Y_2\|_F^2 \tag{8}$$

is a Grassmann kernel.

Proof The kernel is well-defined because $k_P(Y_1, Y_2) = k_P(Y_1R_1, Y_2R_2)$ for any $R_1, R_2 \in \mathcal{O}(m)$. The positive definiteness follows from the properties of the Frobenius norm. For all $Y_1, ..., Y_n (Y_i \in \mathcal{G})$ and $c_1, ..., c_n (c_i \in \mathbb{R})$ for any $n \in \mathbb{N}$, we have

$$\begin{aligned} \sum_{ij} c_i c_j \|Y_i'Y_j\|_F^2 &= \sum_{ij} c_i c_j \text{tr}(Y_iY_i'Y_jY_j') \\ = \text{tr}(\sum_i c_i Y_iY_i')^2 &= \|\sum_i c_i Y_iY_i'\|_F^2 \geq 0. \quad \blacksquare \end{aligned}$$

We can generate a family of kernels from the Projection kernel. For example, the square-root $\|Y_i'Y_j\|_F$ is also a positive definite kernel.

[2] The metric properties follow from the properties of matrix 2-norm and F-norm. To check the conditions in Def. 3 for Procrustes we use the equality $\min_{R_1,R_2} \|Y_1R_1 - Y_2R_2\|_{2,F} = \min_{R_3} \|Y_1 - Y_2R_3\|_{2,F}$ for $R_1, R_2, R_3 \in \mathcal{O}(m)$.

4.2. Binet-Cauchy Metric

The Binet-Cauchy metric can also be understood from an embedding. Let s be a subset of $\{1, ..., D\}$ with m elements $s = \{r_1, ..., r_m\}$, and $Y^{(s)}$ be the $m \times m$ matrix whose rows are the $r_1, \ ... \ , r_m$-th rows of Y. If $s_1, s_2, ..., s_n$ are all such choices of the subset s ordered lexicographically, then the Binet-Cauchy embedding is defined as

$$\Psi_{BC} : \mathcal{G}(m, D) \to \mathbb{R}^n, \quad Y \mapsto \left(\det Y^{(s_1)}, ..., \det Y^{(s_n)} \right), \tag{9}$$

where $n = {}_D C_m$ is the number of choosing m rows out of D rows. The natural innerproduct in this case is $\sum_{r=1}^{n} \det Y_1^{(s_i)} \det Y_2^{(s_i)}$.

Proposition 2 *The Binet-Cauchy kernel*

$$k_{BC}(Y_1, Y_2) = (\det Y_1' Y_2)^2 = \det Y_1' Y_2 Y_2' Y_1 \tag{10}$$

is a Grassmann kernel.

Proof First, the kernel is well-defined because $k_{BC}(Y_1, Y_2) = k_{BC}(Y_1 R_1, Y_2 R_2)$ for any $R_1, R_2 \in \mathcal{O}(m)$. To show that k_{BC} is positive definite it suffices to show that $k(Y_1, Y_2) = \det Y_1' Y_2$ is positive definite. From the Binet-Cauchy identity, we have

$$\det Y_1' Y_2 = \sum_s \det Y_1^{(s)} \det Y_2^{(s)}.$$

Therefore, for all $Y_1, ..., Y_n (Y_i \in \mathcal{G})$ and $c_1, ..., c_n (c_i \in \mathbb{R})$ for any $n \in \mathbb{N}$, we have

$$\begin{aligned} \sum_{ij} c_i c_j \det Y_i' Y_j &= \sum_{ij} c_i c_j \sum_s \det Y_i^{(s)} \det Y_j^{(s)} \\ &= \sum_s \left(\sum_i c_i \det Y_i^{(s)} \right)^2 \geq 0. \end{aligned}$$ ∎

We can also generate another family of kernels from the Binet-Cauchy kernel. Note that although $\det Y_1' Y_2$ is a Grassmann kernel we prefer using $k_{BC}(Y_1, Y_2) = \det(Y_1' Y_2)^2$, since it is directly related to principal angles $\det(Y_1' Y_2)^2 = \prod \cos^2 \theta_i$, whereas $\det Y_1' Y_2 \neq \prod \cos \theta_i$ in general.[3] Another variant $\arcsin k_{BC}(Y_1, Y_2)$ is also a positive definite kernel[4] and its induced metric $d = (\arccos(\det Y_1' Y_2))^{1/2}$ is a conditionally positive definite metric.

4.3. Indefinite Kernels from Other Metrics

Since the Projection metric and the Binet-Cauchy metric are derived from positive definite kernels, all the kernel-based algorithms for Hilbert spaces are at our disposal. In contrast, other metrics in the previous sections are not associated with any Grassmann kernel. To show this we can use the following result (Schoenberg, 1938; Hein et al., 2005):

Proposition 3 *A metric d is induced from a positive definite kernel if and only if*

$$\hat{k}(x_1, x_2) = -d^2(x_1, x_2)/2, \quad x_1, x_2 \in \mathcal{X} \tag{11}$$

is conditionally positive definite.

The proposition allows us to show a metric's non-positive definiteness by constructing an indefinite kernel matrix from (11) as a counterexample.

There have been efforts to use indefinite kernels for learning (Ong et al., 2004; Haasdonk, 2005), and several heuristics have been proposed to make an indefinite kernel matrix to a positive definite matrix (Pekalska et al., 2002). However, we do not advocate the use of the heuristics since they change the geometry of the original data.

5. Grassmann Discriminant Analysis

In this section we give an example of the Discriminant Analysis on Grassmann space by using kernel LDA with the Grassmann kernels.

5.1. Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) (Fukunaga, 1990), followed by a K-NN classifier, has been successfully used for classification.

Let $\{\boldsymbol{x_1}, ..., \boldsymbol{x_N}\}$ be the data vectors and $\{y_1, ..., y_N\}$ be the class labels $y_i \in \{1, ..., C\}$. Without loss of generality we assume the data are ordered according to the class labels: $1 = y_1 \leq y_2 \leq ... \leq y_N = C$. Each class c has N_c number of samples.

Let $\boldsymbol{\mu_c} = 1/N_c \sum_{\{i|y_i=c\}} \boldsymbol{x_i}$ be the mean of class c, and $\boldsymbol{\mu} = 1/N \sum_i \boldsymbol{x_i}$ be the overall mean. LDA searches for the discriminant direction $\boldsymbol{w}$ which maximizes the Rayleigh quotient $L(\boldsymbol{w}) = \boldsymbol{w}' S_b \boldsymbol{w} / \boldsymbol{w}' S_w \boldsymbol{w}$ where S_b and S_w are the between-class and within-class covariance matrices respectively:

$$\begin{aligned} S_b &= \frac{1}{N} \sum_{c=1}^{C} N_c (\boldsymbol{\mu_c} - \boldsymbol{\mu})(\boldsymbol{\mu_c} - \boldsymbol{\mu})' \\ S_w &= \frac{1}{N} \sum_{c=1}^{C} \sum_{\{i|y_i=c\}} (\boldsymbol{x_i} - \boldsymbol{\mu_c})(\boldsymbol{x_i} - \boldsymbol{\mu_c})' \end{aligned}$$

The optimal $\boldsymbol{w}$ is obtained from the largest eigenvector of $S_w^{-1} S_b$. Since $S_w^{-1} S_b$ has rank $C - 1$, there are

[3] $\det Y_1' Y_2$ can be negative whereas $\prod \cos \theta_i$, the product of singular values, is nonnegative by definition.

[4] Theorem 4.18 and 4.19 (Schölkopf & Smola, 2001).

$C-1$-number of local optima $W = \{\boldsymbol{w_1}, ..., \boldsymbol{w_{C-1}}\}$. By projecting data onto the space spanned by W, we achieve dimensionality reduction and feature extraction of data onto the most discriminant subspace.

5.2. Kernel LDA with Grassmann Kernels

Kernel LDA can be formulated by using the kernel trick as follows. Let $\boldsymbol{\phi} : \mathcal{G} \to \mathcal{H}$ be the feature map, and $\Phi = [\boldsymbol{\phi_1} ... \boldsymbol{\phi_N}]$ be the feature matrix of the training points. Assuming w is a linear combination of the those feature vectors, $\boldsymbol{w} = \Phi\boldsymbol{\alpha}$, we can rewrite the Rayleigh quotient in terms of $\boldsymbol{\alpha}$ as

$$L(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}'\Phi' S_B \Phi\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\Phi' S_W \Phi\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}' K(V - \mathbf{1}_N\mathbf{1}_N'/N)K\boldsymbol{\alpha}}{\boldsymbol{\alpha}'(K(I_N - V)K + \sigma^2 I_N)\boldsymbol{\alpha}}, \tag{12}$$

where K is the kernel matrix, $\mathbf{1}_N$ is a uniform vector $[1 \ ... \ 1]'$ of length N, V is a block-diagonal matrix whose c-th block is the uniform matrix $\mathbf{1}_{N_c}\mathbf{1}_{N_c}'/N_c$, and $\sigma^2 I_N$ is a regularizer for making the computation stable. Similarly to LDA, the set of optimal $\boldsymbol{\alpha}$'s are computed from the eigenvectors.

The procedures for using kernel LDA with the Grassmann kernels are summarized below:

Assume the D by m orthonormal bases $\{Y_i\}$ are already computed from the SVD of sets in the data.

Training:

1. *Compute the matrix $[K_{\text{train}}]_{ij} = k_P(Y_i, Y_j)$ or $k_{BC}(Y_i, Y_j)$ for all Y_i, Y_j in the training set.*
2. *Solve $\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha})$ by eigen-decomposition.*
3. *Compute the $(C-1)$-dimensional coefficients $F_{\text{train}} = \boldsymbol{\alpha}' K_{\text{train}}$.*

Testing:

1. *Compute the matrix $[K_{\text{test}}]_{ij} = k_P(Y_i, Y_j)$ or $k_{BC}(Y_i, Y_j)$ for all Y_i in training set and Y_j in the test set.*
2. *Compute the $(C-1)$-dim coefficients $F_{\text{test}} = \boldsymbol{\alpha}' K_{\text{test}}$.*
3. *Perform 1-NN classification from the Euclidean distance between F_{train} and F_{test}.*

Another way of applying LDA to subspaces is to use the Projection embedding Ψ_P (7) or the Binet-Cauchy embedding Ψ_{BC} (9) directly. A subspace is represented by a D by D matrix in the former, or by a vector of length $n = {}_D C_m$ in the latter. However, using these embeddings to compute S_b or S_w is a waste of computation and storage resources when D is large.

5.3. Other Subspace-Based Algorithms

5.3.1. Mutual Subspace Method (MSM)

The original MSM (Yamaguchi et al., 1998) performs simple 1-NN classification with d_{Max} with no feature extraction. The method can be extended to any distance described in the paper. There are attempts to use kernels for MSM (Sakano, 2000). However, the kernel is used only to represent data in the original space, and the algorithm is still a 1-NN classification.

5.3.2. Constrained MSM

Constrained MSM (Fukui & Yamaguchi, 2003) is a technique that applies dimensionality reduction to bases of the subspaces in the original space. Let $G = \sum_i Y_i Y_i'$ be the sum of the projection matrices and $\{\boldsymbol{v_1}, ..., \boldsymbol{v_D}\}$ be the eigenvectors corresponding to the eigenvalues $\{\lambda_1 \leq ... \leq \lambda_D\}$ of G. The authors claim that the first few eigenvectors $\boldsymbol{v_1}, ..., \boldsymbol{v_d}$ of G are more discriminative than the later eigenvectors, and they suggest projecting the basis vectors of each subspace Y_1 onto the span$(\boldsymbol{v_1}, ..., \boldsymbol{v_l})$, followed by normalization and orthonormalization. However these procedure lack justifications, as well as a clear criterion for choosing the dimension d, on which the result crucially depends from our experience.

5.3.3. Discriminant Analysis of Canonical Correlations (DCC)

DCC (Kim et al., 2007) can be understood as a nonparametric version of linear discrimination analysis using the Procrustes metric (6). The algorithm finds the discriminating direction $\boldsymbol{w}$ which maximize the ratio $L(\boldsymbol{w}) = \boldsymbol{w}' S_B \boldsymbol{w} / \boldsymbol{w}' S_w \boldsymbol{w}$, where S_b and S_w are the nonparametric between-class and within-class 'covariance' matrices:

$$\begin{aligned} S_b &= \sum_i \sum_{j \in B_i} (Y_i U - Y_j V)(Y_i U - Y_j V)' \\ S_w &= \sum_i \sum_{j \in W_i} (Y_i U - Y_j V)(Y_i U - Y_j V)', \end{aligned}$$

where U and V are from (1). Recall that $\text{tr}(Y_i U - Y_j V)(Y_i U - Y_j V)' = \|Y_i U - Y_j V\|_F^2$ is the squared Procrustes metric. However, unlike our method, S_b and S_w do not admit a geometric interpretation as true covariance matrices, and cannot be kernelized either. A main disadvantage of the DCC is that the algorithm iterates the two stages of 1) maximizing the ratio $L(\mathbf{w})$ and of 2) computing S_b and S_w, which results in computational overheads and more parame-

ters to be determined. This reflects the complication of treating the problem in a Euclidean space with a non-Euclidean distance.

6. Experiments

In this section we test the Grassmann Discriminant Analysis for 1) a face recognition task and 2) an object categorization task with real image databases.

6.1. Algorithms

We use the following six methods for feature extraction together with an 1-NN classifier.

1) GDA1 (with Projection kernel), 2) GDA2 (with Binet-Cauchy kernel), 3) Min dist , 4) MSM, 5) cMSM, and 6) DCC.

For GDA1 and GDA2, the optimal values of σ are found by scanning through a range of values. The results do not seem to vary much as long as σ is small enough. The Min dist is a simple pairwise distance which is not subspace-based. If Y_i and Y_j are two sets of basis vectors: $Y_i = \{\boldsymbol{y_{i1}}, ..., \boldsymbol{y_{im_i}}\}$ and $Y_j = \{\boldsymbol{y_{j1}}, ..., \boldsymbol{y_{jm_j}}\}$, then $d_{\mathrm{Mindist}}(Y_i, Y_j) = \min_{k,l} \|\boldsymbol{y_{ik}} - \boldsymbol{y_{jl}}\|_2$. For cMSM and DCC, the optimal dimension l is found by exhaustive searching. For DCC, we have used two nearest-neighbors for B_i and W_i in Sec. 5.3.3. Since the S_w and S_b are likely to be rank deficient, we first reduced the dimension of the data to $N - C$ using PCA as recommended. Each optimization is iterated 5 times.

6.2. Testing Illumination-Invariance with Yale Face Database

The Yale face database and the Extended Yale face database (Georghiades et al., 2001) together consist of pictures of 38 subjects with 9 different poses and 45 different lighting conditions. Face regions were cropped from the original pictures, resized to 24×21 pixels ($D = 504$), and normalized to have the same variance. For each subject and each pose, we model the illumination variations by a subspace of the size $m = 1, ..., 5$, spanned by the 1 to 5 largest eigenvectors from SVD. We evaluate the recognition rate of subjects with nine-fold cross validation, holding out one pose of all subjects from the training set and using it for test.

The recognition rates are shown in Fig. 2. The GDA1 outperforms the other methods consistently. The GDA2 also performs well for small m, but performs worse as m becomes large. The rates of the others also seem to decrease as m increases. An interpretation of the observation is that the first few eigenvectors from the data already have enough information and the smaller eigenvectors are spurious for discriminating the subjects.

6.3. Testing Pose-Invariance with ETH-80 Database

The ETH-80 (Leibe & Schiele, 2003) database consists of pictures of 8 object categories ('apple', 'pear', 'tomato', 'cow', 'dog', 'horse', 'cup', 'car'). Each category has 10 objects that belong to the category, and each object is recorded under 41 different poses. Images were resized to 32×32 pixels ($D = 1024$) and normalized to have the same variance. For each category and each object, we model the pose variations by a subspace of the size $m = 1, ..., 5$, spanned by the 1 to 5 largest eigenvectors from SVD. We evaluate the classification rate of the categories with ten-fold cross validation, holding out one object instance of each category from the training set and using it for test.

The recognition rates are also summarized in Fig. 2. The GDA1 also outperforms the other methods most of the time, but the cMSM performs better than GDA2 as m increases. The rates seem to peak around $m = 4$ and then decrease as m increases. This results is consistent with the observation that the eigenvalues from this database decrease more gradually than the eigenvalues from the Yale face database.

7. Conclusion

In this paper we have proposed a Grassmann framework for problem in which data consist of subspaces. By using the Projection metric and the Binet-Cauchy metric, which are derived from the Grassmann kernels, we were able to apply kernel methods such as kernel LDA to subspace data. In addition to having theoretically sound grounds, the proposed method also outperformed state-of-the-art methods in two experiments with real data. As a future work, we are pursuing a better understanding of probabilistic distributions on the Grassmann manifold.

References

Absil, P., Mahony, R., & Sepulchre, R. (2004). Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Appl. Math.*, *80*, 199–220.

Chang, J.-M., Beveridge, J. R., Draper, B. A., Kirby, M., Kley, H., & Peterson, C. (2006). Illumination face spaces are idiosyncratic. *IPCV* (pp. 390–396).

Chikuse, Y. (2003). *Statistics on special manifolds, lecture notes in statistics, vol. 174.* New York: Springer.

Edelman, A., Arias, T. A., & Smith, S. T. (1999). The

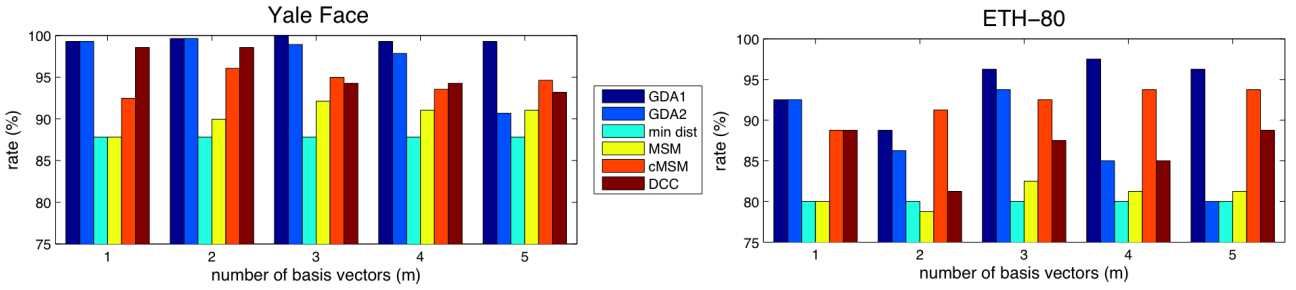

Figure 2. Recognition rates of subjects from Yale face database (Left), and classification rates of categories in ETH-80 database (Right). The bars represent the rates of six algorithms (GDA1, GDA2, Min Dist, MSM, cMSM, DCC) evaluated for $m = 1, ..., 5$ where m is the number of basis vectors for subspaces. The GDA1 achieves the best rates consistently, and the GDA2 also performs competitively for small m.

geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, *20*, 303–353.

Faragó, A., Linder, T., & Lugosi, G. (1993). Fast nearest-neighbor search in dissimilarity spaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, *15*, 957–962.

Fukui, K., & Yamaguchi, O. (2003). Face recognition using multi-viewpoint patterns for robot vision. *Int. Symp. of Robotics Res.* (pp. 192–201).

Fukunaga, K. (1990). *Introduction to statistical pattern recognition (2nd ed.)*. San Diego, CA, USA: Academic Press Professional, Inc.

Georghiades, A. S., Belhumeur, P. N., & Kriegman, D. J. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, *23*, 643–660.

Golub, G. H., & Loan, C. F. V. (1996). *Matrix computations (3rd ed.)*. Baltimore, MD, USA: Johns Hopkins University Press.

Haasdonk, B. (2005). Feature space interpretation of svms with indefinite kernels. *IEEE Trans. Pattern Anal. Mach. Intell.*, *27*, 482–492.

Hein, M., Bousquet, O., & Schölkopf, B. (2005). Maximal margin classification for metric spaces. *J. Comput. Syst. Sci.*, *71*, 333–359.

Kim, T.-K., Kittler, J., & Cipolla, R. (2007). Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. Pattern Anal. Mach. Intell.*, *29*, 1005–1018.

Kondor, R. I., & Jebara, T. (2003). A kernel between sets of vectors. *Proc. of the 20th Int. Conf. on Mach. Learn.* (pp. 361–368).

Leibe, B., & Schiele, B. (2003). Analyzing appearance and contour based methods for object categorization. *CVPR*, *02*, 409.

Ong, C. S., Mary, X., Canu, S., & Smola, A. J. (2004). Learning with non-positive kernels. *Proc. of 21st Int. Conf. on Mach. Learn.* (p. 81). New York, NY, USA: ACM.

Pekalska, E., Paclik, P., & Duin, R. P. W. (2002). A generalized kernel approach to dissimilarity-based classification. *J. Mach. Learn. Res.*, *2*, 175–211.

Sakano, H.; Mukawa, N. (2000). Kernel mutual subspace method for robust facial image recognition. *Proc. of Int. Conf. on Knowledge-Based Intell. Eng. Sys. and App. Tech.* (pp. 245–248).

Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.*, *44*, 522–536.

Schölkopf, B., & Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA, USA: MIT Press.

Shakhnarovich, G., John W. Fisher, I., & Darrell, T. (2002). Face recognition from long-term observations. *Proc. of the 7th Euro. Conf. on Computer Vision* (pp. 851–868). London, UK.

Turk, M., & Pentland, A. P. (1991). Eigenfaces for recognition. *J. Cog. Neurosc.*, *3*, 71–86.

Vishwanathan, S., & Smola, A. J. (2004). Binet-cauchy kernels. *Proc. of Neural Info. Proc. Sys.*.

Wang, L., Wang, X., & Feng, J. (2006). Subspace distance analysis with application to adaptive bayesian algorithm for face recognition. *Pattern Recogn.*, *39*, 456–464.

Wolf, L., & Shashua, A. (2003). Learning over sets using kernel principal angles. *J. Mach. Learn. Res.*, *4*, 913–931.

Wong, Y.-C. (1967). Differential geometry of Grassmann manifolds. *Proc. of the Nat. Acad. of Sci., Vol. 57*, 589–594.

Yamaguchi, O., Fukui, K., & Maeda, K. (1998). Face recognition using temporal image sequence. *Proc. of the 3rd. Int. Conf. on Face & Gesture Recognition* (p. 318). Washington, DC, USA: IEEE Computer Society.

Zhou, S. K., & Chellappa, R. (2006). From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space. *IEEE Trans. Pattern Anal. Mach. Intell.*, *28*, 917–929.